Accuracy of Speech Recognition in Oral Reading Fluency for Diverse Student Groups

Joseph F. T. Nese
Akihito Kamata

February, 2020
Poster presented at the Council for Exceptional Children annual meeting

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140203 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Accuracy of Speech Recognition in Oral Reading Fluency for Diverse Student Groups

Joseph F. T. Nese [1]
jnese@uoregon.edu
Akihito Kamata [2]

[1] Behavioral Research and Teaching, University of Oregon
[2] Center on Research and Evaluation, Southern Methodist University

INTRODUCTION

Automatic speech recognition (**ASR**) can be used to score oral reading fluency (**ORF**) assessments to ameliorate current inadequacies (e.g., administration errors, high opportunity cost), and represents an important part of a larger solution to improve traditional ORF. But more research is needed on how ASR performs for diverse student groups.

The purpose of this study is to examine the accuracy of **ORF** scores as generated by **ASR** compared to humans, and in particular, differential effects for students with disabilities (**SWD**) and those receiving English language (**EL**) supports.

Research Questions

1. Are the agreement rates of ORF word scores between the humans and ASR lower for SWDs or ELs?
2. Are the differences in ORF WCPM between humans and ASR exacerbated for SWDs or ELs?

Sample. The total sample size was N = 650 students.

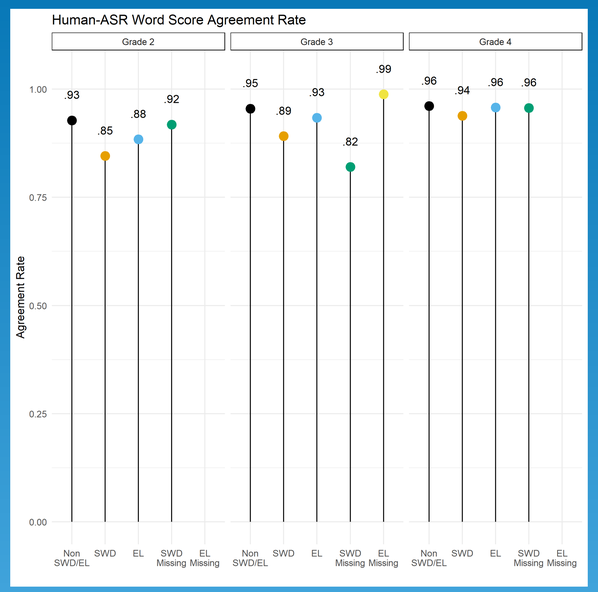
Characteristic[1]	Grade 2, N = 153	Grade 3, N = 182	Grade 4, N = 315
Sex			
Female	67 (44%)	79 (43%)	116 (37%)
Male	73 (48%)	64 (35%)	118 (37%)
Missing	13 (8.5%)	39 (21%)	81 (26%)
Ethnicity			
Hispanic/Latino	28 (18%)	26 (14%)	41 (13%)
Not Hispanic/Latino	112 (73%)	117 (64%)	193 (61%)
Missing	13 (8.5%)	39 (21%)	81 (26%)
Students with a Disability (SWD)			
Yes	21 (14%)	11 (6.0%)	33 (10%)
No	119 (78%)	132 (73%)	201 (64%)
Missing	13 (8.5%)	39 (21%)	81 (26%)
English Learners (EL)			
Yes	17 (11%)	12 (6.6%)	17 (5.4%)
No	123 (80%)	131 (72%)	217 (69%)
Missing	13 (8.5%)	39 (21%)	81 (26%)

[1] Statistics presented: n (%)

METHOD

The following R (R Core Team 2019) packages were used: Arnold (2019); Iannone, Cheng, and Schloerke (2020); Sjoberg et al. (2019); Bates et al. (2015); Teh (2015); Xie, Lesur, and Thorne (2019); Thorne (2019); Allaire et al. (2019); Chan et al. (2018); Wickham et al. (2019).

Automatic speech recogntion was less accurate scoring words for SWDs... but the difference in scoring was mitigated when reading scores were aggregated for passages.



RESULTS

RQ 1: We fit mixed-effect generalized linear models (GLM) for each grade with random effects for *student* and *passage*, and regressed the word score agreement rate (proportion of words scored correct or incorrect by both the human and ASR for each student reading) on *disability* and *EL* status.

Across Grades 2 to 4, **the ORF word score agreement rates between human criterion and ASR were significantly lower for SWDs compared to their non-SWD/non-EL peers**. There was no such difference for EL students .

Results of Word Score Agreement Rate Mixed-Effect GLMs, by Grade

	Grade 2				Grade 3				Grade 4			
	Estimate	SE	z-value	p-value	Estimate	SE	z-value	p-value	Estimate	SE	z-value	p-value
Fixed Effects												
Intercept[1]	2.55	0.08	30.86	> .001	3.05	0.07	43.89	> .001	3.20	0.09	34.99	> .001
SWD-Missing	−0.14	0.23	−0.60	.551	−1.53	0.98	−1.56	.118	−0.11	0.08	−1.41	.159
SWD	−0.85	0.20	−4.26	> .001	−0.94	0.19	−4.85	> .001	−0.48	0.14	−3.39	> .001
EL-Missing	–	–	–	–	1.37	0.97	1.41	.159	–	–	–	–
EL	−0.52	0.22	−2.41	.016	−0.40	0.19	−2.10	.035	−0.09	0.25	−0.35	.725
Random Effects[2]												
Passages	0.25	–	–	–	0.26	–	–	–	0.66	–	–	–
Students	1.04	–	–	–	0.93	–	–	–	0.98	–	–	–

[1] The intercept represents non-SWD and non-EL students.
[2] Estimates reflect the standard deviations of the random effects.

RQ 2: We fit mixed-effect linear models for each grade with random effects for *student* and *passage*, and regressed WCPM difference score (the human criterion score minus the ASR score) on *disability* and *EL* status.

The differences in ORF WCPM scores between human and ASR were not exacerbated for SWD or EL students.

	Grade 2			Grade 3			Grade 4		
	Estimate	SE	t-value	Estimate	SE	t-value	Estimate	SE	t-value
Fixed Effects									
Intercept[1]	4.52	0.83	5.42	3.96	0.59	6.74	4.76	0.71	6.73
SWD-Missing	−0.99	2.36	−0.42	4.06	8.62	0.47	−0.84	1.01	−0.83
SWD	−1.02	2.06	−0.50	0.24	1.72	0.14	−0.85	1.43	−0.59
EL-Missing	–	–	–	−4.25	8.53	−0.50	–	–	–
EL	−2.30	2.22	−1.04	1.48	1.67	0.89	−2.08	2.26	−0.92
Random Effects[2]									
Passages	2.21	–	–	1.64	–	–	3.39	–	–
Students	10.59	–	–	8.04	–	–	8.68	–	–
Residual	8.10	–	–	8.28	–	–	8.40	–	–

[1] The intercept represents non-SWD and non-EL students.
[2] Estimates reflect the standard deviations of the random effects.

Conclusion. We speculate that the ASR may be less accurate than a human scorer for SWDs at the word level, but the difference in scoring is mitigated when scores are aggregated at the passage level.

REFERENCES

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2019. *Rmarkdown: Dynamic Documents for R*. https://CRAN.R-project.org/package=rmarkdown.

Arnold, Jeffrey B. 2019. *Ggthemes: Extra Themes, Scales and Geoms for 'Ggplot2'*. https://CRAN.R-project.org/package=ggthemes.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Chan, Chung-hong, Geoffrey CH Chan, Thomas J. Leeper, and Jason Becker. 2018. *Rio: A Swiss-Army Knife for Data File I/O*.

Iannone, Richard, Joe Cheng, and Barret Schloerke. 2020. *Gt: Easily Create Presentation-Ready Display Tables*. https://github.com/rstudio/gt.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sjoberg, Daniel D., Margie Hannum, Karissa Whiting, and Emily C. Zabor. 2019. *Gtsummary: Presentation-Ready Data Summary and Analytic Result Tables*. https://CRAN.R-project.org/package=gtsummary.

Teh, Victor. 2015. *Qrcode: QRcode Generator for R*. https://CRAN.R-project.org/package=qrcode.

Thorne, Brent. 2019. *Posterdown: Generate Pdf Conference Posters Using R Markdown*. https://CRAN.R-project.org/package=posterdown.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui, Romain Lesur, and Brent Thorne. 2019. *Pagedown: Paginate the Html Output of R Markdown with Css for Print*. https://CRAN.R-project.org/package=pagedown.

UNIVERSITY OF OREGON